Induction of recurrent break cluster genes in neural progenitor cells differentiated from embryonic stem cells in culture

Aseda Tena[a,b,c], Yuxiang Zhang[a,b,d], Nia Kyritsis[a,b], Anne Devorak[a,b], Jeffrey Zurita[a,b], Pei-Chi Wei[a,b,1,2], and Frederick W. Alt[a,b,d,1]

[a]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA 02115; [b]Department of Genetics, Harvard Medical School, Boston, MA 02115; [c]Biological and Biomedical Sciences PhD Program, Harvard University, Boston, MA 02115; and [d]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115

Contributed by Frederick W. Alt, March 17, 2020 (sent for review December 20, 2019; reviewed by Jayanta Chaudhuri and Andre Nussenzweig)

Mild replication stress enhances appearance of dozens of robust recurrent genomic break clusters, termed RDCs, in cultured primary mouse neural stem and progenitor cells (NSPCs). Robust RDCs occur within genes ("RDC-genes") that are long and have roles in neural cell communications and/or have been implicated in neuropsychiatric diseases or cancer. We sought to develop an in vitro approach to determine whether specific RDC formation is associated with neural development. For this purpose, we adapted a system to induce neural progenitor cell (NPC) development from mouse embryonic stem cell (ESC) lines deficient for XRCC4 plus p53, a genotype that enhances DNA double-strand break (DSB) persistence to enhance detection. We tested for RDCs by our genome-wide DSB identification approach that captures DSBs via their ability to join to specific genomic Cas9/single-guide RNA–generated bait DSBs. In XRCC4/p53-deficient ESCs, we detected seven RDCs, all of which were in genes and two of which were robust. In contrast, in NPCs derived from these ESC lines we detected 29 RDCs, a large fraction of which were robust and associated with long, transcribed neural genes that were also robust RDC-genes in primary NSPCs. These studies suggest that many RDCs present in NSPCs are developmentally influenced to occur in this cell type and indicate that induced development of NPCs from ESCs provides an approach to rapidly elucidate mechanistic aspects of NPC RDC formation.

recurrent DNA break cluster genes | embryonic stem cells | differentiated neural progenitor cells | transcription

The DNA ligase 4, and its obligate XRCC4 cofactor, are core classical nonhomologous end-joining (C-NHEJ) factors that are required for mouse lymphocyte development, due to essential roles in V(D)J recombination (1, 2). Moreover, each of them is also required for neural development (2, 3). Absence of either of these factors led to widespread p53-dependent death of newly generated neurons due to the inability to repair double-strand breaks (DSBs) generated in neuroprogenitors (4, 5). While p53 deficiency rescued XRCC4-deficient neuronal apoptosis, the p53/XRCC4 double-deficient mice routinely died from medulloblastomas with recurrent genomic rearrangements characteristic of those found in the sonic hedgehog subgroup of this human childhood brain cancer (6, 7). The striking overlaps of these findings with those obtained for the same mutant genotypes with respect to effects on development of lymphocytes and cancers of progenitor lymphocytes led to speculation that specific DSBs may impact neural development or disease (8–10). However, identification of such putative breaks was challenging with available technologies. To elucidate potential recurrent DSBs in developing neural progenitors, we employed linear amplification-mediated, high-throughput, genome-wide, translocation sequencing (LAM-HTGTS) to map, at nucleotide resolution, DSBs in primary XRCC4/p53-deficient neural stem and progenitor cells (NSPCs) based on ability to join to bait DSBs introduced by Cas9/single-guide RNAs (sgRNAs) (11). Our initial LAM-HTGTS experiments on ex vivo propagated mouse NSPCs employed HTGTS bait DSBs on three different mouse chromosomes and identified 27 recurrent DSB clusters that, strikingly, were nearly all in bodies of long neural genes and mostly only observed after treatment with aphidicolin (APH) to induce mild replication stress (11).

To extend these studies, we exploited our finding that joining of two DSBs occurs more frequently if they lie on the same *cis* chromosome (12, 13). Thus, we introduced DSBs into each of the 20 mouse chromosomes as baits for HTGTS libraries from control or APH-treated NSPCs (14). This analysis confirmed previously identified recurrent DNA break clusters (RDCs) and identified many more. Again, most RDCs were in genes and were identified upon NSPC treatment with APH to generate mild replication stress. Based on identification frequency with different baits and RDC-DSB density, we ranked relative RDC-gene robustness. On this basis, 19 originally identified plus 11

Significance

We previously discovered a set of long neural genes susceptible to frequent DNA breaks in primary mouse brain progenitor cells. We termed these genes RDC-genes. RDC-gene breakage during brain development might alter neural gene function and contribute to neurological diseases and brain cancer. To provide an approach to characterize the unknown mechanism of neural RDC-gene breakage, we asked whether RDC-genes appear in neural progenitors differentiated from embryonic stem cells in culture. Indeed, robust RDC-genes appeared in neural progenitors differentiated in culture and many overlapped with robust RDC-genes in primary brain progenitors. These studies indicate that in vitro development of neural progenitors provides a model system for elucidating how RDC-genes are formed.

Author contributions: A.T., P.-C.W., and F.W.A. designed research; A.T., Y.Z., A.D., and P.-C.W. performed research; A.T., Y.Z., N.K., J.Z., P.-C.W., and F.W.A. analyzed data; and A.T., P.-C.W., and F.W.A. wrote the paper.

Reviewers: J.C., Memorial Sloan Kettering Cancer Center; and A.N., National Cancer Institute.

The authors declare no competing interest.

Data deposition: Sequencing data used for analyses and figure preparation presented in this study have been deposited in the Gene Expression Omnibus database, https://www.ncbi.nlm.nih.gov/geo/ (accession no. GSE 142315).

[1]To whom correspondence may be addressed. Email: p.wei@dkfz-heidelberg.de or alt@enders.tch.harvard.edu.

[2]Present address: Functional and Structural Genomics, German Cancer Research Center, 69120 Heidelberg, Germany.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1922299117/-/DCSupplemental.

First published April 24, 2020.

newly identified RDC-genes were highly robust (14). Of note, four of these highly robust RDC-genes were identifiable in untreated controls but became more robust with APH treatment, consistent with ectopic replication stress augmenting an ongoing endogenous process (11, 14). The great majority of highly robust RDC-genes are very long (>0.5 Mb), variably transcribed, encode proteins that regulate synaptic function/cell adhesion, and have been associated with neuropsychiatric and development disorders and/or cancer (11, 14). We categorized such genes as group 1 RDCs. We also identified group 2 RDCs that contain multiple genes with at least one greater than 80 kb long, and group 3 RDCs which are clusters of small (<20 kb) genes. Group 2 and 3 RDCs are mostly less robust than group 1 RDCs and also less frequently associated with neuropsychiatric and neurodevelopmental disorders (14) and are not a focus of this study. Recently, LAM-HTGTS studies identified 36 DSB clusters in very long genes (analogous to group 1 RDCs) in human neural precursors cells derived from human-induced pluripotent stem cells, of which about 70% were orthologs of mouse RDC-genes (15).

Some RDCs overlap with common fragile sites (CFSs) and copy number variations (CNVs), which have been suggested to be fragile due to collisions between transcription and replication related processes in very long genes (11, 16–18). However, detailed mechanisms that cause RDC-gene DSBs in neural progenitor cells (NPCs) largely remain to be elucidated. For example, given the variable levels of overall robust RDC-gene transcription (11, 14), the precise role of transcription in RDC generation, or even if it is required, is an important, open question. Likewise, the question of when robust RDCs arise in the context of NSPC development and the factors, including stress, that promote their occurrence is also unknown. Addressing such questions experimentally in primary NSPCs that arise in vivo during mouse development is very challenging technically. Therefore, to facilitate mechanistic and developmental studies of RDC-gene formation in NSPCs, we sought to employ an in vitro system for induction of NPC differentiation from embryonic stem cells (ESCs) (19). Comparison of RDC-genes in ESCs to those in NPCs derived from them in vitro (ESC-NPCs) could provide insights into mechanisms that lead to robust RDC occurrence during NPC development. Likewise, if RDCs arise de novo in such an in vitro differentiation system, targeted genetic modifications could be introduced to RDC-genes or their regulatory sequences in parental ESCs and assess effects on RDC formation subsequent to their induction to ESC-NPC or even to mature neurons in culture. Here, we report that robust RDC formation does indeed occur in ESC-NPCs generated form ESCs in culture.

Results

HTGTS Bait DSBs to Identify RDCs in ESCs. In our initial RDC experiments, we identified RDCs in primary NSPCs via introduction of Cas9/sgRNA-mediated bait DSBs into specific sites on three mouse chromosomes, including chromosomes (chr) 12, 15, and 16 (11). To investigate if mouse ESCs also harbor RDCs, we used the same general strategy, with chrs. 12 and 15 HTGTS baits as in the earlier experiments, along with a chr7 HTGTS bait (*SI Appendix*, Table S1), as the chr16 HTGTS bait did not yield robust cutting in ESCs (*Discussion*). The Cas9/sgRNA constructs were individually introduced in *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESCs (referred to as ESC line 1) on one chromosomal site at a time. Based on our previous studies, the *Xrcc4*$^{-/-}$*p53*$^{-/-}$ background leads to DSB persistence, which similarly to primary NSPC studies can facilitate detection of ESC RDCs (11, 14). In each experiment, ESCs were treated with APH to induce mild replication stress or with dimethyl sulfoxide (DMSO) (vehicle control). Experiments with each bait DSB were repeated four times and analyzed as described (11). For HTGTS library data analyses, we applied our custom-designed RDC pipeline (20), which identified five RDCs in the first tested ESC line (ESC line 1), which only appeared after APH treatment (Fig. 1 *A–C*). These five RDCs were all located within genes (Fig. 1 *D–H*). To corroborate these findings, we repeated the experiments in a second genotype-matched ESC line (referred to as ESC line 2) and identified five APH-induced RDCs (*SI Appendix*, Fig. S1 *A–C*) that partially overlapped (three of five) with the set in ESC line 1 (*SI Appendix*, Fig. S1 *D–H*).

The seven confirmed ESC RDCs all occurred within group 1 RDC-genes (Fig. 1 *D–H* and *SI Appendix*, Fig. S1 *D–H*) and were among the longest, transcribed ESC genes, with the majority being late-replicating (*SI Appendix*, Fig. S2*A* and Table S4). Three ESC RDC-genes (*Auts2*, *Wwox*, and *Fhit*) were shared between both lines, but only *Auts2* and *Wwox* were classified as robust RDCs based on high RDC DSB junction density and the fact that they were captured by all three different HTGTS baits (Fig. 1 *D* and *E* and *SI Appendix*, Fig. S1 *D* and *E*). *Fhit*, along with the four ESC RDCs that were unique to each line, were considered weaker RDCs (lower DSB junctions and discovered by only two of the three baits) (Fig. 1 *F–H* and *SI Appendix*, Fig. S1 *F–H*). Notably, the two unique RDCs in the ESC line 1 (*Gpc6* and *Bai3*) were both RDC candidates in ESC line 2 (i.e., found with one HTGTS bait), while *Dock1*, a unique RDC in ESC line 2, was a candidate in ESC line 1 (Dataset S2). We note that the majority of RDCs identified in both ESC lines (six of seven) were identified in *trans* by DSB baits employed on chr12, chr15, and chr7 (Fig. 1 *D–H* and *SI Appendix*, Fig. S1 *D–F* and *H*). Only *Dock1,* an RDC-gene in ESC line 2, resided on a bait chromosome (*SI Appendix*, Fig. S1*G*). We also note that four RDCs (*Auts2, Gpc6, Wwox*, *and Fhit*) overlapped with previously reported CNVs in ESCs (17), supporting the notion that RDC DSBs may contribute to formation of CNVs in *Xrcc4*$^{-/-}$*p53*$^{-/-}$ *ESC* lines (21). Furthermore, these RDCs have been mapped as CFSs in human fibroblast lines (17, 22, 23). Finally, all seven RDCs identified in ESCs were identified as RDCs in prior NSPC studies (11, 14) and four (*Auts2*, *Wwox*, *Gpc6*, and *Bai3*) were also RDC-genes in ESC-NPCs (discussed below). *Dock1* was an RDC candidate in both ESC-NPC lines (Dataset S3).

Identification of RDCs in ESC-Induced NPCs. We differentiated both *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC line 1 and 2 into *Xrcc4*$^{-/-}$*p53*$^{-/-}$ NPCs to test whether this process would lead to generation of NPCs that had formed additional RDCs compared to those in their parental ESCs. For both *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC lines, we employed immunofluorescence assays to confirm the differentiation of *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESCs into NPCs based on lack of expression of an ESC-specific *Oct4* and gain of expression of the NPC-specific *Sox1* and *Nestin* markers (*SI Appendix*, Fig. S3*A*). We further tested *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC-derived NPCs (*Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC-NPCs) by measuring their ability to give rise to neurons in a monolayer differentiation culture system (19). After 10 d in culture medium supplemented with retinoic acid, *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC-NPCs down-regulated *Sox1* and gained expression of neuronal-specific markers *NeuN* and *β-III tubulin* (*SI Appendix*, Fig. S3*B*). To test the ability of *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC-NPCs to generate RDCs, we performed HTGTS based on Cas9/sgRNA-specific bait DSBs introduced into chr12, chr15, and chr7, as described above for ESC HTGTS experiments. After 12 d of differentiation, ESC-NPCs were treated with APH or DMSO and nucleofected, and HTGTS experiments were performed 4 d later. Experiments were repeated four times and analyzed as previously described (11, 14). For *Xrcc4*$^{-/-}$*p53*$^{-/-}$ NPCs derived from ESC line 1, we identified 22 RDCs (Dataset S1), which, as for primary NSPC RDCs, were enhanced by replication stress (Fig. 2 *A–C*) and located within genes (Fig. 2 *D–G*). We also performed a second set of studies with *Xrcc4*$^{-/-}$*p53*$^{-/-}$ NPCs derived from ESC line 2, which identified 24 RDCs (Dataset S1),

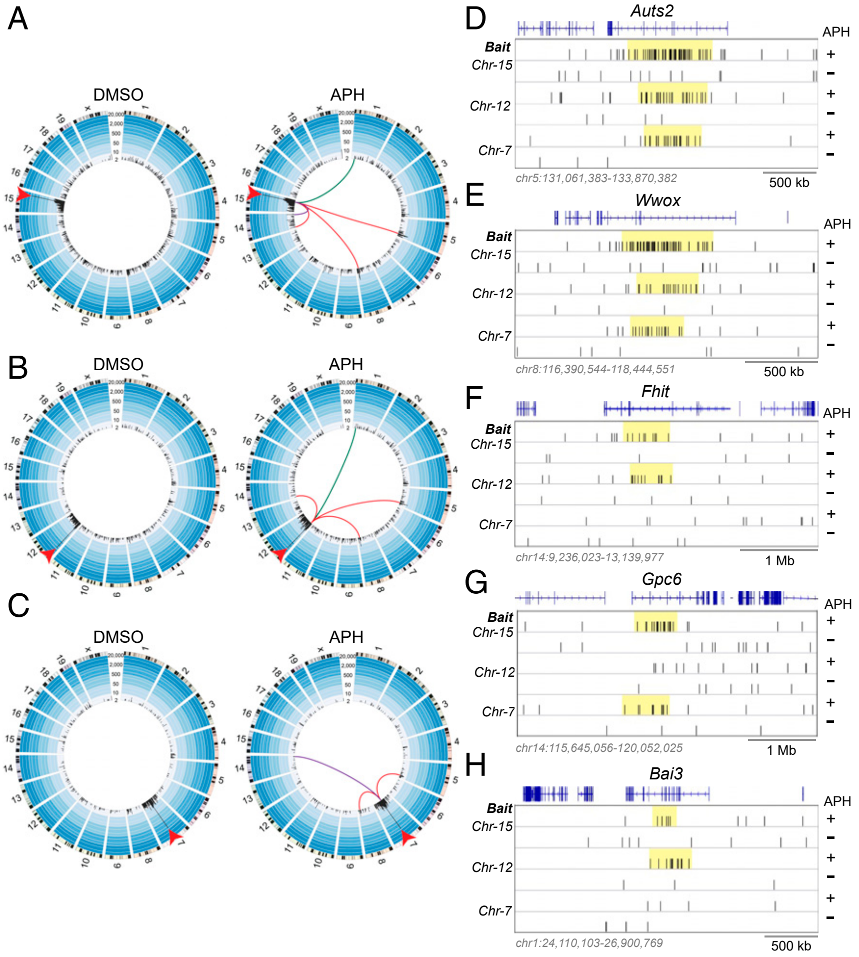

Fig. 1. Genome-wide Identification of replication stress-induced RDC-genes in ESCs. (*A*–*C*) Circos plots of the mouse genome divided into individual chromosomes show the genome-wide LAM-HTGTS junction pattern in *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESCs. Junctions identified by LAM-HTGTS baits locating at chr15, chr12, or chr7 were shown as black bars in per 2.5-Mb bin. Bar height indicates the number of translocations per bin on a log scale. Ten thousand randomly selected junctions from four independent experiments are plotted in DMSO- (*Left*) and APH-treated cells (*Right*). Red arrowheads in circos plots denote the bait DSB site for each bait chromosome. Red lines in APH-treated experiments (*Right*) connect the break site to three replication stress-induced RDCs identified by bait DSBs on all three tested chromosomes. The purple and green lines connect the break site to RDCs identified by two of the three HTGTS baits. (*D*–*H*) Twenty thousand randomly selected LAM-HTGTS prey junctions from APH- (+) or DMSO-treated (–) ESCs are plotted. Panels represent five RDC-genes discovered by either two or three independent HTGTS baits on the indicated chromosomes. The yellow rectangle indicates the RDC location. RefGene (blue track) indicates the gene location.

which were again enhanced by replication stress (*SI Appendix*, Fig. S3 *C*–*E*) and located in genes (*SI Appendix*, Fig. S3 *F*–*I*). In both ESC-NPC lines we detected the same off-targets (OTs) from chr7 and chr12 sgRNAs (Fig. 2 *B* and *C* and *SI Appendix*, Fig. S3 *D* and *E*) as in NSPCs (11, 14). Chr15 sgRNA OTs were also detected in both ESC-NPC lines (Fig. 2*A* and *SI Appendix*, Fig. S3*A*) but were different from those previously found (11) because we used a different chr15 sgRNA in these experiments (*SI Appendix*, *Material and Methods* and Tables S1 and S2). Notably, we did not detect any OT sites for all of the sgRNAs in ESC lines (Fig. 1 *A*–*C* and *SI Appendix*, Fig. S1 *A*–*C*).

Between both *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC-NPC cell lines, we identified 29 RDC-genes, of which 17 appeared in *Xrcc4*$^{-/-}$*p53*$^{-/-}$ NPCs derived from both parental *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC lines (Fig. 2 *D*–*G* and *SI Appendix*, Figs. S3 *F*–*I* and S4 *A* and *B*). In addition, five RDCs were found only in ESC line 1-derived *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC-NPCs and seven RDCs were found only in ESC line 2-derived *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC-NPCs (*SI Appendix*, Fig. S4 *C* and *D*). Notably, three of five NPC line 1-specific RDCs were RDC candidates in NPC line 2, and five out of seven NPC line 2-specific RDCs were RDC candidates in NPC line 1 (Dataset S3). Among these RDC-genes, 27 were group 1 RDCs, ranking among the longest, transcribed ESC-NPC genes, with the majority being late-replicating (*SI Appendix*, Fig. S2*B* and Table S4). The other two RDCs were group 2 (*Tpgs2*/*Celf4*), and group 3 (*Ackr2*/*1700048O20Rik*) RDCs did not rank among the longest genes and were not late-replicating (*SI Appendix*, Fig. S2*B* and Table S4). Comparative analysis of the 29 ESC-NPC RDCs to those previously identified in primary NSPCs with chr12, chr15, and chr7 HTGTS baits revealed that the majority overlapped between ESC-NPCs and NSPCs, with a subset being found only in NSPCs or ESC-NPCs (*SI Appendix*, Table S5). All NSPC-specific RDCs were RDC candidate genes in ESC-NPCs (Dataset S3). Similar to RDCs identified in primary NSPCs (11, 14), DSBs in ESC-NPCs RDC-genes map broadly across the length of the RDC-gene transcription unit, (Fig. 2 *D*–*G* and *SI Appendix*, Figs. S3 *F*–*I* and S4 *A*–*D*). Of the 17 shared RDCs, 14 were highly robust in both ESC-NPCs lines (Fig. 3*A*, *SI Appendix*, Fig. S5*A*, and Dataset S1) and among the 30 most robust RDCs-genes detected in primary NSPCs by employing baits from all chromosomes (14). These findings demonstrate that *Xrcc4*$^{-/-}$*p53*$^{-/-}$ NPCs differentiated from *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESCs in culture

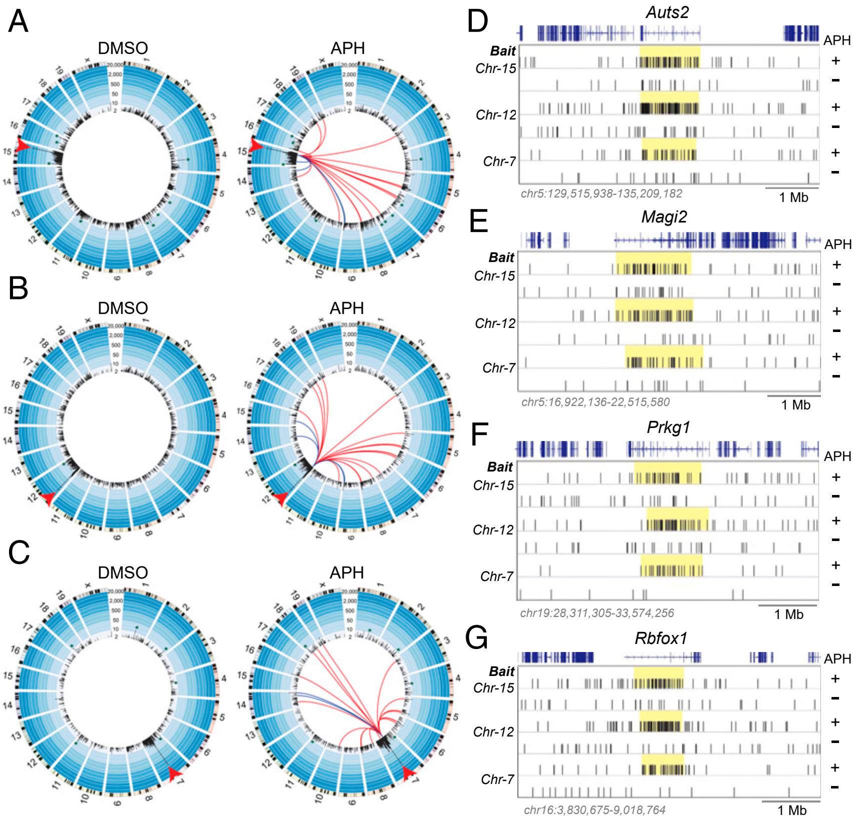

Fig. 2. Genome-wide identification of replication stress-induced RDC-genes in ESC-derived NPCs. (*A*–*C*) Circos plots of the mouse genome divided into individual chromosomes show the genome-wide LAM-HTGTS junction pattern of chr15, chr12, and chr7 sgRNA-mediated bait DSBs in $Xrcc4^{-/-}p53^{-/-}$ ESC-NPCs. Plots are organized as described in Fig. 1 *A*–*C*. Ten thousand randomly selected junctions from four independent experiments are plotted for DMSO- (*Left*) and APH-treated (*Right*) for each LAM-HTGTS bait experiment. Red lines in APH-treated experiments (*Right*) connect the break site to 17 replication stress-induced RDCs identified by bait DSBs on all three tested chromosomes. Blue lines connect break site to RDCs identified by bait DSBs on two of the three tested break sites, which numbered four for chr15, four for chr12, and two for chr7 bait DSBs. LAM-HTGTS bait DSB sites (red arrowhead) and Cas9:sgRNA OT sites (green asterisk) are denoted in both DMSO and APH plots. (*D*–*G*) Twenty thousand randomly selected LAM-HTGTS prey junctions from APH- (+) or DMSO-treated (–) experiments are plotted for *Auts2, Magi2, Prkg1*, and *Rbfox1* RDC-genes. The yellow rectangle indicates the RDC location. RefGene (blue track) indicates the gene location.

gain the ability to generate a robust subset of RDCs observed in $Xrcc4^{-/-}p53^{-/-}$ primary NSPCs. We also identified many RDC candidates in both ESCs and ESC-NPCs based on junctions detected with one bait, with the majority of these weak RDC candidates residing on a bait-containing chromosome (Datasets S2 and S3). It is possible that additional ESC and ESC-NPC RDCs would be confirmed if baits from all chromosomes were used for analyses (11, 14).

Transcription Activity of RDC-Genes in ESCs and NPCs. Transcription patterns of RDC-genes found in $Xrcc4^{-/-}p53^{-/-}$ ESC lines versus RDC-genes found in $Xrcc4^{-/-}p53^{-/-}$ ESC-NPCs derived from them might provide clues to mechanisms underlying RDC-gene formation. We analyzed the transcription of all RDC-genes identified in $Xrcc4^{-/-}p53^{-/-}$ ESCs and $Xrcc4^{-/-}p53^{-/-}$ ESC-NPCs experiments via global run-on sequencing (GRO-seq). The 7 RDC-genes identified in $Xrcc4^{-/-}p53^{-/-}$ ESCs and the 29 RDCs found in $Xrcc4^{-/-}p53^{-/-}$ ESC-NPCs showed no consistent relationship between transcription levels and RDC formation. Some of the RDC-genes had high transcriptional activity in both cell types, others had high transcription activity only in $Xrcc4^{-/-}p53^{-/-}$ ESC-NPCs, while others had low transcription in $Xrcc4^{-/-}p53^{-/-}$ESCs and $Xrcc4^{-/-}p53^{-/-}$ ESC-NPCs (Fig. 3*B* and *SI Appendix*, Fig. S5*B*). Moreover, the 14 robust $Xrcc4^{-/-}p53^{-/-}$ ESC-NPC RDCs also lacked a strong correlation between transcription levels and RDC-gene robustness score (Fig. 3 *A* and *B* and *SI Appendix*, Fig. S5 *A* and *B*). More specifically, *Auts2* and *Wwox* are robust RDC-genes in both $Xrcc4^{-/-}p53^{-/-}$ ESCs and $Xrcc4^{-/-}p53^{-/-}$ ESC-NPCs and have high transcription activity in both cell types (Fig. 3*B*, left box; Fig. 3 *C* and *F*); *Dcc* and *Ctnna2* are $Xrcc4^{-/-}p53^{-/-}$ ESC-NPC–specific RDCs with high transcription activity only in $Xrcc4^{-/-}p53^{-/-}$ ESC-NPCs and low transcription activity in $Xrcc4^{-/-}p53^{-/-}$ ESCs (Fig. 3*B* middle box; Fig. 3 *D* and *G*); while *Csmd1* and *Nrg3* are robust RDCs in ESC-NPCs despite their low transcription activity in these cells (Fig. 3*B* third quadrant; Fig. 3 *E* and *H*). Similar examples are shown for the $Xrcc4^{-/-}p53^{-/-}$ ESC line 2 and its derivative $Xrcc4^{-/-}p53^{-/-}$ ESC-NPCs (*SI Appendix*, Fig. S5 *A*, *B*, and *C*–*H*). Thus, these ESC and ESC-NPC studies, together with our primary NSPCs studies (11, 14), indicate that transcription levels of RDC-genes per se have no obvious correlation with RDC formation in them.

Discussion

There are many fundamental questions related to molecular mechanisms that give rise to DSBs within RDC-genes in primary NSPCs that remain unanswered, in part due to difficulty in studying this phenomenon in primary cells from mice. We now show that induced development of NPCs from ESCs in vitro leads to the formation of a robust set of NPC-specific RDC-genes, many of which overlap with robust RDC-genes in primary NSPCs. Therefore, the in vitro NPC differentiation system

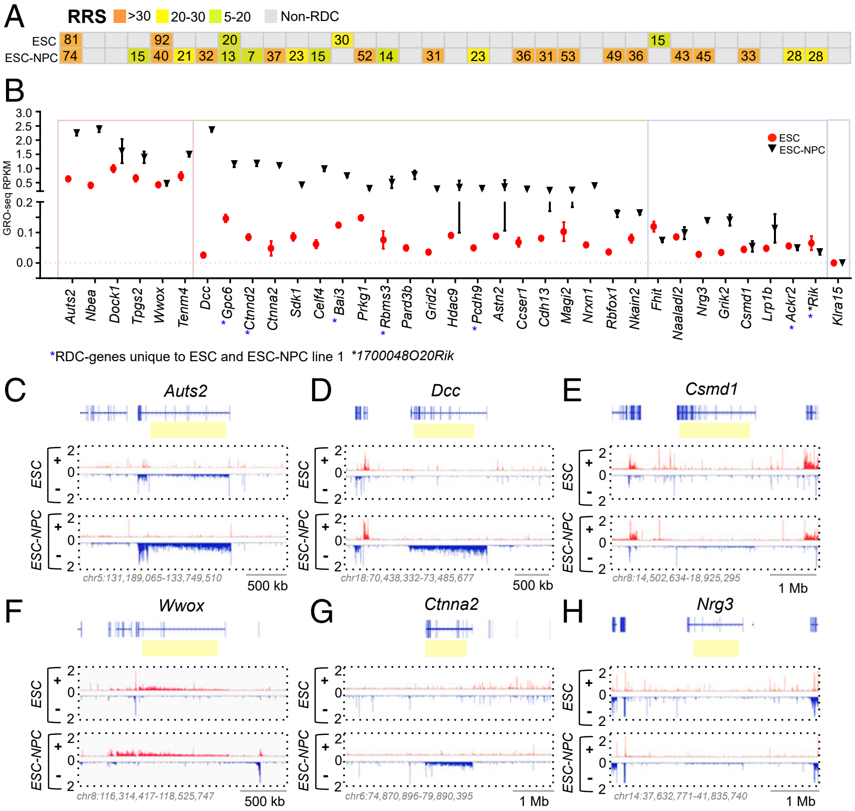

Fig. 3. Transcription activity in ESCs and NPCs. (*A*) RDC RRS is indicated for each RDC-gene. Differences in RRS values (highest–lowest) are color-coded as illustrated. (*B*) Transcription activity measured by reads per kilobase per million (RPKM) from GRO-seq experiments is plotted for RDC-genes in both ESCs and differentiated NPCs. The leftmost box in the graph contains RDC-genes that have higher transcription activity in both ESCs and ESC-NPCs. The middle box contains RDC-genes that have higher transcription activity only in ESC-NPCs. The second-to-right box contains genes with lower transcription activity in both cell types, and the rightmost box contains *Klra15,* which is not an RDC-gene and has no transcription activity in both cell types (RPKM = 0). *Klra15* is used to demonstrate a no-transcription level in relation to low transcription levels. Data represent (mean ± SD) based on three repeats. Blue asterisk: RDC-genes that are unique to cell line 1. (*C*–*H*) Robust RDC-genes in ESCs and differentiated NPCs with different transcription levels in both cell lines. Ordinate indicates normalized GRO-seq counts. Blue color indicates transcription orientation in centromeric-to-telomeric direction and red color telomeric-to-centromeric direction. RefGene (blue track) indicates the gene location and the yellow rectangle highlights the RDC location.

will offer a rapid approach to target genetic modifications in ESCs and then test their effects on robust RDC formation in ESC-NPCs. In some cases, mutations of interest could further be tested by differentiating the ESC-NPC cultures to neurons (19) or by using the ESCs for neural blastocyst complementation in vivo studies (24). The ESC-NPC differentiation studies showed that a large number of robust RDCs are absent in ESCs and only appear upon differentiation of ESCs into NPCs. Moreover, 26 of 36 replication-stress susceptible DSB "hotspot" genes identified in human-induced pluripotent stem cell-derived neural precursor cells were orthologs of primary mouse NSPC RDC-genes (11, 14), and 12 were orthologs of robust ESC-NPC RDCs identified in this study (15). In this regard, further studies with the ESC-NPC differentiation system to elucidate mechanisms underlying RDC formation should also be relevant to RDC formation in human NPCs.

Our current studies address, but do not fully answer, the important question of whether the apparently NPC-specific RDCs observed in ESC-NPCs versus their ESC progenitors result from gain of a neuronal program or loss of an ESC program. Thus, our current studies are consistent with one or the other possible mechanisms, or a combination of both. Some apparent differences in RDCs and OTs observed between the two cell types could reflect differential Cas9/sgRNA cutting at bait sites in the two different cell types. In addition, dominant homologous recombination (HR) repair pathways in ESCs (25, 26) might obviate some (but not all) RDC DSBs that could otherwise contribute to translocations in ESCs by decreasing the frequency of persistent RDC (or bait) DSBs in ESCs, thereby leading to fewer translocations between bait DSBs and RDC DSBs. As HR occurs during the S and G2 cell-cycle phases when templates presented by sister chromatids are available (27), it theoretically may contribute substantially to repair of ESC RDC DSBs given the potential roles of DNA replication in RDC formation (11, 14, 15). In somatic cells, such as NPCs, (C-NHEJ, which can occur throughout the cell cycle, is thought to be more dominant (26). In this regard, our studies are all done in C-NHEJ-deficient cells, which means that in ESCs and ESC-NPCs, RDC breaks repaired by end joining must employ alternative end joining, which generally is considered less efficient than C-NHEJ. However, we also note that RDCs are observed in wild-type NSPCs (11) and that C-NHEJ deficiency is thought to only increase their detection by HTGTS by allowing DSBs to persist longer and thus more robustly contribute to translocations. Again, such a DSB persistence effect may not be as prominent in ESCs if HR repairs a larger fraction of their DSBs than it does in ESC-NPCs. Finally, on the other hand, it remains possible that transcription has qualitative roles in ESC-NPCs RDC formation, not necessarily related to absolute levels, that preferentially activate some RDC DSBs in ESC-NPCs versus ESCs.

Our studies confirm that the majority of highly robust NSPC-RDCs and, now, ESC-NPC RDC-genes tend to arise in very long, transcribed neural genes associated with specific brain functions (refs. 11 and 14 and this study), a phenomenon that is also characteristic of human NPC RDCs (15). Early studies performed on mouse ESCs and human fibroblasts found both spontaneous and replication-induced CNVs which were linked to long genes that were actively transcribed and late-replicating (17, 21). A majority of mouse hotspot CNVs were also CFSs in human fibroblasts (22, 23). These studies led to speculation that mechanisms of breakage and CNV formation may involve collisions between the replication and transcription machinery that would be accentuated due to the long replication time of long genes (16–18, 28). Furthermore, because a subset of the robust primary NSPC and ESC-NPCs RDC-genes overlapped with mouse ESC CNVs and human fibroblast CFSs, it was proposed that this mechanism could potentially apply to RDC-gene formation (11, 14, 15, 17). Yet, both our prior NSPC and current ESC-NPC RDC studies indicate that there is no direct relationship between robust RDC occurrence and the relative level of RDC-gene transcription. However, it is possible that very low transcription is sufficient for robust RDC formation. We can now directly test this notion by disrupting transcription of both highly and weakly transcribed robust RDC-genes in the in vitro NPC differentiation system and, if warranted by results, employ additional gene-targeted mutational approaches to further test the potential roles of transcription. Studies could also be readily envisioned to employ the NPC in vitro differentiation approach to test other factors related to robust ESC-NPC RDC formation including potential roles of gene length, replication timing, and replication stress.

Materials and Methods

Cell Culture and LAM-HTGTS Bait DSB Induction. We used two genotype-matched, de novo-derived ESCs that were *Xrcc4*$^{-/-}$*p53*$^{-/-}$ for the described studies. ESC-NPCs were generated as described (19) with minor modifications. LAM-HTGTS bait DSB induction at chr7, chr12, and chr15 was performed as described in refs. 11 and 14 and *SI Appendix*. Details of cell cultures and LAM-HTGTS bait induction are provided in *SI Appendix, Materials and Methods*.

LAM-HTGTS. We prepared, sequenced, processed, aligned, and analyzed the LAM-HTGTS libraries as described (11, 14, 20), except that reads were also sequenced on Next-Seq. *SI Appendix*, Table S3 lists the number of unique junctions used for RDC identification in each experiment.

RDC Identification. A SICER-based, unbiased, genome-wide method and a MACS-based method were both applied to identify APH-induced RDCs in *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESCs and *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC-NPCs as described previously (11). RDC robustness score (RRS) was calculated as previously described in ref. 14. Details are also provided in *SI Appendix, Materials and Methods*.

GRO-Seq. GRO-seq libraries were prepared as previously described (29). Three experimental replicates were performed for both *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC lines and *Xrcc4*$^{-/-}$*p53*$^{-/-}$ ESC-NPC lines. Libraries were sequenced on Illumina Hi-Seq and Next-Seq. Details are provided in *SI Appendix, Materials and Methods*.

Data Availability. All of the sequencing data used for analyses and figure preparation presented in this study have been deposited and are available for downloading in the Gene Expression Omnibus database (accession no. GSE 142315).

ACKNOWLEDGMENTS. This work was supported by a Children's Hospital Department of Medicine and Charles H. Hood Foundation grant. P.-C.W. was supported by Charles A. King Trust Postdoctoral Research Fellowship Program, Bank of America, N.A., Co-Trustees. Y.Z. is supported by a Damon Runyon Fellowship Program. F.W.A. is an Investigator of the Howard Hughes Medical Institute.

1. K.-M. Frank *et al*., Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. *Nature* **396**, 173–177 (1998).
2. Y. Gao *et al*., A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis. *Cell* **95**, 891–902 (1998).
3. D.-E. Barnes, G. Stamp, I. Rosewell, A. Denzel, T. Lindahl, Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. *Curr. Biol.* **8**, 1395–1398 (1998).
4. K.-M. Frank *et al*., DNA ligase IV deficiency in mice leads to defective neurogenesis and embryonic lethality via the p53 pathway. *Mol. Cell* **5**, 993–1002 (2000).
5. Y. Gao *et al*., Interplay of p53 and DNA-repair protein XRCC4 in tumorigenesis, genomic stability and development. *Nature* **404**, 897–900 (2000).
6. C.-T. Yan *et al*., XRCC4 suppresses medulloblastomas with recurrent translocations in p53-deficient mice. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7378–7383 (2006).
7. M. Ratnaparkhe *et al*., Defective DNA damage repair leads to frequent catastrophic genomic events in murine and human tumors. *Nat. Commun.* **9**, 4760 (2018).
8. E.-C. Gilmore, R. S. Nowakowski, V. S. Caviness, Jr, K. Herrup, Cell birth, cell death, cell diversity and DNA breaks: How do they all fit together? *Trends Neurosci.* **23**, 100–105 (2000).
9. M. J. McConnell *et al*.; Brain Somatic Mosaicism Network, Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science* **356**, eaal1641 (2017).
10. I.-L. Weissman, F. H. Gage, A mechanism for somatic brain mosaicism. *Cell* **164**, 593–595 (2016).
11. P.-C. Wei *et al*., Long neural genes harbor recurrent DNA break clusters in neural stem/progenitor cells. *Cell* **164**, 644–655 (2016).
12. Y. Zhang *et al*., Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908–921 (2012).
13. F.-W. Alt, Y. Zhang, F. L. Meng, C. Guo, B. Schwer, Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell* **152**, 417–429 (2013).
14. P.-C. Wei *et al*., Three classes of recurrent DNA break clusters in brain progenitors identified by 3D proximity-based break joining assay. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 1919–1924 (2018).
15. M. Wang *et al*., Increased neural progenitor proliferation in a hiPSC model of autism induces replication stress-associated genome instability. *Cell Stem Cell* **26**, 221–233.e6 (2020).
16. A. Helmrich, M. Ballarino, L. Tora, Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell* **44**, 966–977 (2011).
17. T. E. Wilson *et al*., Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res.* **25**, 189–200 (2015).
18. T. W. Glover, T. E. Wilson, Molecular biology: Breaks in the brain. *Nature* **532**, 46–47 (2016).
19. Q.-L. Ying, M. Stavridis, D. Griffiths, M. Li, A. Smith, Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat. Biotechnol.* **21**, 183–186 (2003).
20. J. Hu *et al*., Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nat. Protoc.* **11**, 853–871 (2016).
21. M.-F. Arlt, S. Rajendran, S. R. Birkeland, T. E. Wilson, T. W. Glover, De novo CNV formation in mouse embryonic stem cells occurs in the absence of Xrcc4-dependent nonhomologous end joining. *PLoS Genet.* **8**, e1002981 (2012).
22. M.-F. Arlt *et al*., Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. *Am. J. Hum. Genet.* **84**, 339–350 (2009).
23. B. Le Tallec *et al*., Molecular profiling of common fragile sites in human fibroblasts. *Nat. Struct. Mol. Biol.* **18**, 1421–1423 (2011).
24. A.-N. Chang *et al*., Neural blastocyst complementation enables mouse forebrain organogenesis. *Nature* **563**, 126–130 (2018).
25. A.-N. Pierce, P. Hu, M. Han, N. Ellis, M. Jasin, Ku DNA end-binding protein modulates homologous repair of double-strand breaks in mamalian cells. *Genes Dev.* **15**, 3237–3242 (2001).
26. E.-D. Tichy *et al*., Mouse embryonic stem cells, but not somatic cells, predominantly use homologous recombination to repair double-strand DNA breaks. *Stem Cells Dev.* **19**, 1699–1711 (2010).
27. R. Scully, A. Panday, R. Elango, N.-A. Willis, DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* **20**, 698–714 (2019).
28. A. Aguilera, T. García-Muse, Causes of genome instability. *Annu. Rev. Genet.* **47**, 1–32 (2013).
29. F.-L. Meng *et al*., Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell* **159**, 1538–1548 (2014).